http://dx.doi.org/10.36785/jaes.111459

A Basic Parts of Speech (POS) Tagset for morphological, syntactic and lexical annotations of Saraiki language

Farrukh Javed Saleemi [a], Muhammad Nabeel Asghar [b,*], Sajid Iqbal [b], Muhammad Umar Chaudhry [c], Muhammad Yasir [d], Sibghat Ullah Bazai [e], Muhammad Qasim Khan [f]

[a] Institute of Southern Punjab Multan, Pakistan
[b] Department of Computer Science, Bahauddin Zakariya University Multan;
[c] AiHawks, Multan 60000, umarch.skku@gmail.com
[d] Department of Computer Science, University of Engineering and Technology Lahore, Faisalabad Campus, Pakistan; muhammadyasir@uet.edu.pk
[e] Cyber Security Lab, School of Natural and Computational Sciences, Massey University, Auckland, New Zealand.
[f] SKKU, South Korea

[*]Corresponding Author: nabeel.asghar@bzu.edu.pk

***Abstract*-- One of the important resources required for various Natural Language Processing (NLP) applications like machine translation, information retrieval and text mining, is annotated text corpora. Text corpora annotation process requires parts of speech (POS) tags to mark different parts of text with grammatical annotations in order to identify linguistic properties of a word, sentence or discourse. The process of marking text items is based on two main features 1) grammatical category and 2) context of text (word, sentence or discourse) i.e. relationship with adjacent and related text.Saraiki being one of oldest languages is still resource scarce language in recorded literature as well as in computational context. According to our study, at present, there is no tagset defined for Saraiki language. This work presents first hierarchical POS (MPOST) tag set for the Saraiki language which is designed to be used in morphological, syntactic and lexical annotations of Saraiki language corpora.**

Keywords—Corpora, Parts of Speech (POS); Saraiki; Tag set; Tagging

Date Received: 27-11-2020
Date Accepted: 18-02-2021
Date Published: 08-06-2021

I. INTRODUCTION

Saraiki is an Indo-Aryan language that is spoken mainly in Southern Punjab of Pakistan and is also a minor language in India and resut of the Indian sub-continent. There are around 26 million native language users in Pakistan and India only[1]. It is written in Perso-Arabic script however it has its own set of alphabets that consists of 45 letters. Out of this 45 letter, 39 are same as that of Urdu language and 6 are additional letters. Although some researchers consider it as a dialect of standard Punjabi language however it is a separate language with its own identity [12][24]. There are different dialects of this language that include Multani (Main Saraiki), Thalli (Thal region), Rajanpur (Southern Saraiki), Rohi (Cholistan desert and adjacent areas), Thar (Thar desert and all Sindh region), Majhi and Shahpuri. It is morphologically a rich language with different tones and well-structured sentence architecture.

To process a language computationally, it is required to build its computational resources. Tagset is one of such resources, that is used to mark different syntactic and semantic units of a language. A tag set defines basic entities known as grammatical constructs. Each language that is spoken in world has different categories of words or language units [4][25].

Most common categories include nouns, pronouns, verbs, adverbs, adjectives, proverbs and adjuncts. Each main category can contain multiple sub-categories that in return may contain further sub-categories. This results in hierarchical relationship in tagset. Explicitly assigning a language category to each language unit in given context is known as POS tagging.

There are number of natural language processing (NLP) applications that require annotated corpus. These applications may include text mining, information retrieval, information extraction, machine translation, natural language text generation and text summarization [17][27]. Applying linguistic tags to the text units makes it easy and suitable for machines to understand the text algorithmically. Corpus-driven

[1] (https://en.wikipedia.org/wiki/Saraiki_language)

approaches of NLP heavily depend upon such embedded linguistic information [2] however rule-based and statistical approaches can also produce better results with the use of annotated corpora [28]. Recently neural-network based algorithms have shown remarkable success in solving various artificial intelligence based problems. The annotation process tags text units with linguistic categories like morphological marking, lexical annotation, syntactic constructs and semantic units. Text annotation may be at different levels which depend upon the intended use of corpus however in general deeper annotation yields more information-rich corpus. The richness of linguistic information in corpus depends upon the quality of tagset used for annotation. Saraiki language is one of resource-poor language [14] and according to our study; no tagset exists for this language.

In this work, we introduce a comprehensive tagset for Saraiki language that could be used for POS tagging of Saraiki language corpora. The motivation behind this work is to initiate and support the computational aspects of Saraiki language in order to promote this language and generate NLP applications for the Saraiki speaking community. Rest of the paper is organized as follows: Section 2 lists the background and review of literature about tag sets developed for local languages. Section 3 presents the methodology and validation process used for tag set development. In section 4, the tagset is described with examples. Finally, section 5 presents conclusion and future work.

II. Related Work

The Saraiki language is resource scarce language. Although there is a rich collection of cultural verbal and undocumented heritage, written and recorded resources are very little perhaps scarce. There are few publishers, publishing in this language. Similarly, there are very small number of researchers working for Saraiki Language. A renowned researcher has developed some resources for this language that include Saraiki Qaida, Saraiki Dictionary and Saraiki Grammer [21]. At present, a Saraiki font has been developed and few text editing software are also available. There is little computational research work done in and for this language [9] [15] [22] [26]. The language is also supported by Unicode system now[2].

Figure. 1 shows the alphabets set of Saraiki language.

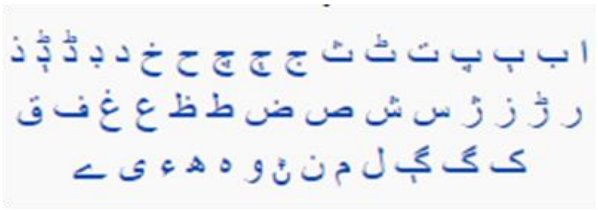

Figure 1: Saraiki Alphabets

Design and development of tagset is the basic step toward building a computational grammar of any language. A tag not only provides the syntactic or morphological category of language unit (token), it also provides whole range of grammatical information for the token. POS tag sets have been developed for most of the languages being spoken in the world. Considering English as an example language, multiple tag sets have been developed [6] [20] [23] [29].

Here we review POS tag set development for Urdu and local languages being spoken in Pakistan and neighboring countries. These local languages are mostly similar to each other however their writing script can be different. Even one local language (i.e. Saraiki) which is same in spoken format however is also written in same script (perso-Arabic). However, the case for Punjabi is not same. For example, in Pakistan, Punjabi is written in Perso-Arabic script (Shahmukhi, پنجابی), however in India it is written in Gurmukhi script (ਪੰਜਾਬੀ) [16].

Considering the Saraiki language, this work only focuses Shahmukhi script tag sets which is based on Persso-Arabic script. The tag sets are normally divided into three categories: flat tag sets, hierarchical tag sets and fine-grained tag sets [30] [18].

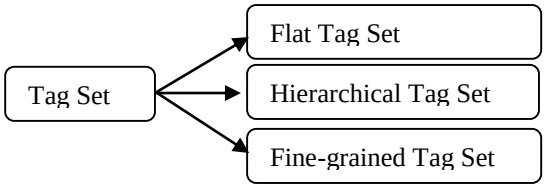

Figure 2: Tag Set Classification

Flat tag sets, provide a tag for each token without any information about relationship among tokens. Although flat tag sets are easy to work with however they fail to capture detailed relationship between linguistic constructs. This results in unsuitability of flat tag-set for various NLP applications that need to capture fine grained details about the corpus.

Instead of designing a large number of independent tags, a small number of master categories are identified and each master level category further contains number of sub-categories. The process of breaking a category into sub-categories and to further sub-sub-categories can be continued to an arbitrary level, depending upon the granularity of required information. This arrangement of linguistic units forms a tree structure. Hierarchical tag sets provide tags as well as information about hierarchical relationship among tags or token categories [31]. Figure 3 shows an example of hierarchical structure of one language unit.

[2] (https://unicode.org/L2/L2002/02274-arabic-ext.pdf).

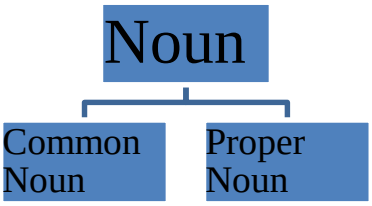

Figure 3: Example Hierarchical Tags

A fine-grained tag set is one which provides very detailed categorization of tokens and tagging scheme for them. It could be considered as deep tree structure [32]. In Pakistan, Urdu is most fortunate language that has computational resources. Baker, Paul [7] produce Urdu annotated corpora for EMILLE project. Hardie [11] developed another Urdu tag set. Next, Sarmad Hussain [13] from CRULP produced another tag set for Urdu and finally they reviewed and improved their own tag set [3][3]. A hierarchical tag set was proposed by [1] to generate KON-TB treebank. For Punjabi language (Shahmukhi script) we could not find any tag set [10]. However, there is a tag set developed for Sindhi language [19]. We tried to find POS tagsets for other local languages however no tag set found so far.

As per our knowledge, Saraiki language has no tagset at present and in this work, as first effort, we present a detailed hierarchical tag set. We first group tokens under main categories and then we devise their sub-categories, relative to each other, forming a tree-like structure.

III. Tagset Design Methodology

As a part of larger work, to annotate Saraiki corpora, to use in Saraiki computational and linguistic research, the development of this tagset is done. During Multani Parts of Speech Tagset (MPOST) development, multiple issues are faced that include unavailability of relevant terms for linguistic units in Saraiki like (lughat). We could not find a standard grammar for Saraiki language and hence we are not able to find terms for nouns, verbs etc (i.e. name of noun in Saraiki). To make an initial effort, most of the terms are borrowed from Urdu. The work was validated by human experts of native language which were the faculty members of department of Saraiki, B.Z.U. Multan. We intend to resolve such issues in next revision of tag-set. Table 1 lists few examples from our manual annotated corpora.

Pseudo Code for Tagset Development

1. Read the word
2. Check its type, i.e. noun, verb or harf
3. If word is noun
   a. Identify its type i.e. proper noun or common noun
      i. If word is personal noun. Classify it among following categories
         1. Noun Personal Common Tool
         2. Noun Personal Common Sound
         3. Noun personal common small
         4. Noun personal common big
         5. Noun Personal Common Empathic Time
      ii. If word is common noun then categorize among followings
         1. Noun Personal Proper Pronoun subjective
         2. Noun Personal Proper Pronoun Objective
         3. Noun Personal Proper Pronoun Possessive
         4. Noun Personal Proper Relative Pronoun
         5. Noun Personal Proper Title
         6. Noun Personal Proper Address
         7. Noun Personal Proper Sur-name
         8. Noun Personal Proper Alias
         9. Noun Personal Proper Title Poetic
         10. Noun Personal Proper Demonstrative
      iii. If word is noun and does not lie in above categories
         1. Noun Adjective Possessive
         2. Noun Adjective Reflexive
         3. Noun Adjective Relative
4. If word is verb then classify it as follows
   a. Past
      i. Near past
      ii. Past supremacy
      iii. Far past
      iv. Past doubt
      v. Past condition
      vi. Past reinforcement
   b. Present
   c. Future
   d. Order
   e. Forbid
   f. Required
   g. Varieties
   h. Verbal verb
5. If word is Harf then classify it as follows
   a. Conjunction Coordinating
   b. Subordinating
   c. Semantic Marker
   d. Key particle
   e. Conjunction
   f. Adjective particle

Table 1: Noun Tag Example in Sentence

Category & tag	Examples

[3] http://www.cle.org.pk/

Noun <NP>	جہاز- زمین – درخت – چھوہر-
آسم	ایہ<NPPP> شہر<NP> ساڈی<NPPG> سنجانڑ<NADJ> ہے <VB>

To validate our work, human experts are consulted. However, the authors believe that there is space for more work in this domain.

IV. Saraiki Hierarchical Tagset

The framework of proposed Saraiki tag set is laid out in a hierarchy of levels and sub-levels. Subclasses are further categories into sub-sub-categories. According to Saraiki Grammar [5] [8] which is lot more aligned with Urdu language, we have identified three main classes:

- **Noun (اسم)** : Like other languages, name of something is considered as Noun. It has two subcategories: Proper Noun (NP) and Noun Adjective (NA) which are further divided among sub-sub-categories. The detailed specification of nouns is given on the pattern of Urdu grammar.
- **Verb (فعل):** This represents a work. Inflectional forms of verb are similar to that as in Urdu language. For example in Urdu sometimes verb inflectional form only can convey the right meaning whereas in some cases, additional words are used to make the sense clear i.e. prefixes, postfixes or infixes are added.
- **Proposition (حرف):** A word that does not possess useful meaning in itself, however when used as connector between other words can clear their meanings.
- **Residual:** These are the words or tokens which could not be placed in any category. Tag **FW** is used to identify this category. For example, if there is word "call" (in Saraiki) or some foreign language word like "Computer", "Masha Allah", these are given FF tag.

Now we list the sub-categories, tags and examples of each of these categories.

A. Noun and its sub-categories

Like Urdu language, nouns are generally inflected for number, case and gender. We present the hierarchical relationship in the form of tables. Our first category is noun which is further divided between two sub-categories. It is to mention that Saraiki language has no well-defined and documented structure.

TABLE 2: NOUN AND ITS ROOT CATEGORIES

Category	Sub-category	Tag	Examples
Noun **(اسم)**	Noun Personal	NP	جہاز- زمین – درخت – چھوہر-
	اسم ذات		
	Noun Adjective اسم صفت	NADJ	**سہنڑاں،کوہڑا،شوم اسلم چنگا< NADJRR> بندہ ہے**

Noun Personal is further sub-divided into two categories: Noun personal common and proper noun. A proper noun is the name given to some specific
entity whereas the common noun is the name given to class of that entity. Noun personal has further sub-categories whereas noun common also has more sub-categories. The details of these categories are given in table 3 and table 4.

TABLE 3: SUB-CATEGORIES OF PERSONAL NOUN

Category	Sub-category		Tag	Examples
Noun Personal Common اسم نکرہ			**NPC**	**لڑکا – عورت – مرد – جوان**
				او ڈاھڈا سوہنڑا جوان <NPC>ہے
	Noun Personal Common Tool اسم آلہ		NPCT	مسواک – ہتھوڑی – کلہاڑی
				لوہار ہتھوڑی < NPCT> نال کم کریندے
	Noun Personal Common Sound اسم صوت		NPCSU	ٹک ٹک، چوں چوں،ٹھا ٹھا
				گھڑی ٹک ٹک<NPCSU> پئ کریندی اے
	Noun Personal Common Small اسم تصغیر		NPCS	نکڑا،نکڑی،بالڑی
				نجمہ نکڑی < NPCS> بالڑی ہے
	Noun Personal Common Big اسم مکبر		NPCB	پگڑا،منجہا
				بابا سر تے پگڑ<NPCB> بدھدی ویندے
	Noun			اج، کل، پرسوں,

	Personal Common Empathic اسم ظرف		NPCEP	اساں پرسوں <NPCEP> آسوں
		Noun Personal Common Empathic Time ظرف زماں	NPCET	فجرے- دیگرے- سونجھلے
				او فجرے > آ <NPCET> ویسی
		Noun Personal Common Empathic ظرف مکاں	NPCEP	
				گھر - مسیت – سکول
				اج سکول > دی <NPCEP> چھٹی ہے

Proper noun is further divided among four sub-categories and in return these sub-categories are further divided among sub-sub-categories. Table 4 lists the sub-categories of proper Noun.

TABLE 4: SUBCATEGORIES OF PROPER NOUN

Category			Tag	Examples
Proper Noun **اسم معرفہ**			**NPP**	**لہور- مُلتان- پاکستان**
				ملتان <NPP> امبا ں دا شہر اے
Noun Personal pronoun اسم ضمیر		NPPP		میں-اساں- تُساں- اِیں-اے- او- اوندا -ایہ
				تساں دا ماما میں <NPPP>ہاں
	Noun Personal Proper Pro-noun subjective ضمیر فاعلی	NPPPS		اوکُوں- تُہاکُوں- انہاں کُوں
				<NPPPS>اُکوں آکھو میں کل نہ آساں
	Noun Personal Proper Pro-Noun Objective	NPPPO		تیکوُں – میکُوں – ساکوں
				اسلم آکھا اے<NPPPO> ساکوں
	ضمیر مفعلئ			
	Noun Personal Proper Pro-noun Possessive ضمیر اصافئ	NPPPP		تُہاڈّا –
				<NPPPP>بھرا کن اے تہاڈّا
Noun Personal Proper Relative Pronoun اسم موصول		NPPR		جو- جہڑّا- جِنہاں
				علئ جہڑّا <NPPR> حامد دا بھرا اے، اوہ آئا ھا
Noun Personal Proper Title اسم علم		NPPT		شاعر مشرق – قایداعظم- موسی کلیم اللہ
				شاعر مشرق<NPPT> کا کلام بھوں اچھا اے
	Noun Personal Proper Address خطاب	NPPA		خان بہادر – رستم زماں
				رستم زماں <NPPA> برصغیر دئ پہچان ھن
	Noun Personal Proper Sur-name کنئت	NPPS		ابن مریم – ابو القاسم
				<NPPS>ابو القاسم بہھوں اچھے ھن
	Noun Personal Proper Alias A عرف	NPPAA		محمود عُرف مودا – نجمہ عُرف نجو
				مودا <NPPAA> بھرا بھون اچھے ھن
	Noun Personal Proper Title Poetic تخلص	NPPTP		غالب- حالی
				غالب <NPPTP>ہیک وڈا شاعر ہای
Noun Personal Proper Demonstrative اسم اشارہ		NPPD		اے – او- خد
				آو <NPPD> اج آسن

Noun adjective is the word that further describes the noun to which it is associated and is subdivided into 3 sub-categories as listed in table 5.

TABLE 5: NOUN ADJECTIVE SUB-CATEGORIES

Category	Tag	Examples
Noun Adjective Possessive صفت ذاتی	NADJP	شریف – چٹا – لال
		میں ہیک شریف < انسان ہاں NADJP>
Noun Adjective Reflexive صفت اصلئ	NADJRR	چنگا - ماڑا
		اسلم چنگا < بندہ ہے NADJRR>
Noun Adjective Relative صفت نسبتئ	NADJA	عربی بندا- مکی / مد نی
		ساڈے شہر وچ عربی بندا<NADJR > آیا ودا ہے

B. *Verb (کم) and its sub-categories*

A uni-gram or multi-gram word that describes some action, occurring or state of something is called Verb. Like Urdu and other local languages, verb in Saraiki is also divided among multiple sub-categories. Table 6 lists verb main category with its tag.

TABLE 6: VERB AND ITS TAG

Entity	Tag	Examples
Verb فعل	VB	کھاونڑاں- ونجنڑاں- کرنا, لکھنڑاں
		اوہ بندے شام کوں گھر ونجندا <VB> ا ہے

There are 9-sub-categories of verb (کم). These sub-categories are formed on the type of work. Table 7 provides details about its sub-categories:

TABLE 7: VERB AND ITS SUB-CATEGORIES

Entity	Tag	Examples
Present فعل حال	VBPR	کھاندا پئے – ویندا پئے
		او روٹی کھاندا<VBPR > پیا ہے
Future فعل مستقبل	VBF	کھا سی – پی سی – مرسی
		او شربت پی سی < VBF>
order فعل امر	VBOR	کرو-اے کم کر
		اپنا کم کرو <VBOR>
forbid فعل نہی	VBFO	نہ بول، مت جا،
		مندا نہ بول <VBFO>
required	VBRQ	ستا- جاگا
		خرگوش بھجا <VBRQ>
فعل لازم		
varieties فعل متعدی	VBVT	خریدناں، لکھناں اور سونا
		اسلم نے خط لکھا <VBVT>
Known فعل معرف	VBKN	کھاندا- گھن آسی
		عمران مجھ گھن آسی <VBKN>
verbal verb فعل مجہول	VBVV	نماز پڑھی- کھایا گیا
		نماز پڑھی <VBVV> ویندی اے
Past فعل ما ضئ	VBPA	گیا – سُتا – کھا دھدا
		ہائ<VBPA >او کراچی گیا

Verb sub-categories given in table 7 can further be divided among sub-sub-categories. Again this classification is made on the properties of verb type.

Table 8 gives further sub-categories of Past tense of Verb.

TABLE 8: PAST TENSE AND ITS SUB-CATEGORIES

Entity	Tag	Examples
Near Past ماضئ قریب	VBPASN	پڑھا اے - لکھا ا ے
		اسلم نے خط لکھا اے<VBPASN>
Past supremacy ماضئ مطلق	VBPAPS	گیا- کھایا
		فرخ لاھور گیا <VBPAPS>
Past dream ماضئ بعید	VBPAPD	لکھا ہائ ، پیتا ہائ
		سلمیٰ نے خط لکھا ہائ <VBPAPD>
Past doubt ماضئ شکیہ	VBPAPD	ہوسی-پوسی –ہوسی آ-
		جاوید ویندا ہوسی <VBPAPD>
past condition or meditation ماضئ تمنائ	VBPACM	آوے ہا- کروں ہا-
		ملتان آوے ھا تا ونجوں ہا <VBPACM>
Past reinforcement ماضئ استمراری	VBPAPR	- اوٹھیندا ہائ-
		عدیل فجرے اوٹھیندا ہائ <VBPAPR>

C. *Harf (≅ Particle):*

In English particle is defined as “a word that can not be inflected” or “a function word associated with other word or phrase to impart meaning”. Saraiki words that lie in this category do not fulfill first definition however second definition can correctly define such words. Following table 9 provides details about Harf and its sub-categories.

TABLE 9: HARF AND ITS SUB-CATEGORIES

Entity	Sub-Entity	Tag	Examples
Particle حرف		**WP**	**ایچ - وچ- ہور**
			اوہ بندے شام کوں گھر ونجندا ا ہے <wp>
	Conjunction Coordinating حرف جار	PWCC	واسطے – اندر – اُتے – تلے
			سکول دے اندر > بچے <PWCC پڑھدے پے ہین
	Subordinating حرف علت	PWCS	جے- کیونکہ – جیکر- تا کہ
			محنت کرو تا کہ > پاس تھی <PWCS ونجوں
	Semantic Marker حرف اضافت	PWCM	دا- دی- دے- نے- وچ- تک-کُول -تِلک –تئن
			بزار ساڈے گھر دے کول<PWCM > ہے
	KEY Particle حرف بیان	PWKP	کے
			سبزی گھن کے > آ <PWKP
	Conjunction حرف عطف	PWCJ	تے – کر – دے
			گھر دے<PWCJ > کم کرنا چنگی گال ہے
	Adjective Particle حرف تشبہ	PWAP	جئیا – جئے- جئ – اونویں
			اسلم جئیے<PWAP>محنتی بندے بہوں گھٹ ہین

V. Conclusion and Future work

In this paper, we have presented a linguistically motivated Parts of Speech tagset, MPOST, for Saraiki language. This hierarchical framework allows designing language-specific tag sets that are interoperable and flexible. The developed tag set could be used as a source for further research on Saraiki NLP, such as design and development of a POS tagger for Saraiki language. The tag set is tested on custom developed corpora using manual annotation.

As our own future work, we intend to develop a MPOST tagger to automatically annotate corpora for Saraiki language.

VI. References

Abbas, Qaiser. "Building a hierarchical annotated corpus of urdu: the URDU. KON-TB treebank." International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, 2012.

Adamou, Evangelia. A corpus-driven approach to language contact: Endangered languages in a comparative perspective. Vol. 12. Walter de Gruyter GmbH & Co KG, 2016.

Ahmed, T., et al. "The CLE urdu POS tagset." poster presentation in Language Resources and Evaluation Conference (LERC 14). 2014.

Ahmad, A.M., Sulong, G., Rehman, A., Alkawaz,MH., Saba, T. (2014) Data Hiding Based on Improved Exploiting Modification Direction Method and Huffman Coding, Journal of Intelligent Systems, vol. 23 (4), pp. 451-459, doi. 10.1515/jisys-2014-0007.

Aman ullah Kazim, "Jamy Saraiki Qwaed", Usman Publications April, 2015, Lahore, Pakistan.

Atwell, E. S., et al. "A comparative evaluation of modern English corpus grammatical annotation schemes." ICAME Journal: International Computer Archive of Modern and Medieval English Journal 24 (2000): 7-23.

Baker, Paul, et al. "EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation." LREC. 2002.

Bashir, Elena et. al. " Grammar of Hindko, Panjabi, and Saraiki", De Gruyter publishers 2018.

Belaïd, Abdel, and Mohamed Imran Razzak. "Middle Eastern Character Recognition." Handbook of Document Image Processing and Recognition. Springer London, 2014. 427-457.

Gill, Mandeep Singh, Gurpreet Singh Lehal, and Shiv Sharma Joshi. "Part of speech tagging for grammar checking of punjabi." The Linguistic Journal 4.1 (2009): 6-21.

Hardie, Andrew. "Developing a tagset for automated part-of-speech tagging in Urdu." Corpus Linguistics 2003. 2003.

Harouni, M., Rahim, M.S.M., Al-Rodhaan,M., Saba, T., Rehman, A., Al-Dhelaan, A. (2014) Online Persian/Arabic script classification without contextual information, The Imaging Science Journal, vol. 62(8), pp. 437-448, doi. 10.1179/1743131X14Y.0000000083.

Hussain, Sarmad. "Resources for urdu language processing." Proceedings of the 6th workshop on Asian Language Resources. 2008.

Hussain, Syed Safdar. "The Growth of Saraiki Language." Pakistan Journal of Social Sciences (PJSS) 36.1 (2016).

Jan, M. T., and Y. Saleem. "Optical Character Recognition (OCR) System For Saraiki Language Using Neural Networks." University of Engineering and Technology Taxila. Technical Journal 21.3 (2016): 106.

Kumar, Sunil (2018). Developing POS Tagset for Dogri. Language in India, 18(1).

Lewinski, Nastassja A., Ivan Jimenez, and Bridget T. McInnes. "An annotated corpus with nanomedicine and pharmacokinetic parameters." International journal of nanomedicine 12 (2017): 7519.

Lung, J.W.J., Salam, M.S.H, Rehman, A., Rahim, M.S.M., Saba, T. (2014) Fuzzy phoneme classification using multi-speaker vocal tract length normalization, IETE Technical Review, vol. 31 (2), pp. 128-136, doi. 10.1080/02564602.2014.892669.

Mahar, Javed Ahmed, and Ghulam Qadir Memon. "Rule based part of speech tagging of sindhi language." Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on. IEEE, 2010.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19.2 (1993): 313-330.

Mughal, Shaukat. "Saraiki qaidah (Saraiki primer)." Multan, Pakistan: Saraiki Isha'ati Idarah.

Nodehi, A. Sulong, G. Al-Rodhaan, M. Al-Dhelaan, A., Rehman, A. Saba, T. (2014) Intelligent fuzzy approach for fast fractal image compression, EURASIP Journal on Advances in Signal Processing, doi. 10.1186/1687-6180-2014-112.

Petrov, Slav, Dipanjan Das, and Ryan McDonald. "A universal part-of-speech tagset." arXiv preprint arXiv:1104.2086 (2011).

Rahman, Tariq. "The Siraiki Movement in Pakistan." Language Problems and Language Planning 19.1 (1995): 1-25.

Rehman, A. Kurniawan, F. Saba, T. (2011) An automatic approach for line detection and removal without smash-up characters, The Imaging Science Journal, vol. 59(3), pp. 177-182, doi. 10.1179/136821910X12863758415649.

Raza, Ghulam. "Reduction of Compound Adpositions in Persian, Urdu and Saraiki." Presentation given at the Sixth International Contrastive Linguistics Conference (ICLC'06), Berlin. Vol. 253. 2010.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In EMNLP, 2015.

Skeppstedt, Maria, et al. "Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study." Journal of biomedical informatics 49 (2014): 148-158.

Younus, Z.S. Mohamad, D. Saba, T. Alkawaz,M.H. Rehman, A. Al-Rodhaan,M. Al-Dhelaan, A. (2015) Content-based image retrieval using PSO and k-means clustering algorithm, Arabian Journal of Geosciences, vol. 8(8) , pp. 6211-6224, doi. 10.1007/s12517-014-1584-7.

Baskaran, Sankaran, et al. "A common parts-of-speech tagset framework for indian languages." In Proc. of LREC 2008. 2008.

Hardie, Andrew. The computational analysis of morphosyntactic categories in Urdu. Diss. Lancaster University, 2004.

Horsmann, Tobias, and Torsten Zesch. "Assigning Fine-grained PoS Tags based on High-precision Coarse-grained Tagging." Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.